German Record Linkage Center
WORKING PAPER SERIES

NO. WP-GRLC-2016-01 | MARCH 30, 2016

Record Linkage of the Linked Employer-Employee Survey of the Socio-Economic Panel Study (SOEP-LEE) and the Establishment History Panel (BHP)

Johanna Eberle | Michael Weinhardt

Record Linkage of the Linked Employer-Employee Survey of the Socio-Economic Panel Study (SOEP-LEE) and the Establishment History Panel (BHP)

Johanna Eberle (Institute for Employment Research)

Michael Weinhardt (University of Bielefeld)

German Record Linkage Center: Working Paper

Contents

Abstract

This working paper describes the linkage of the SOEP-LEE survey of DIW and University of Bielefeld with administrative data on establishments provided by the Institute for Employment Research.

Keywords: Record linkage, administrative data, deterministic matching

1 Introduction

This working paper describes the linkage of the SOEP-LEE survey – which was conducted by TNS infratest on behalf of the German Institute for Economic Research (DIW) in Berlin and the University of Bielefeld – with administrative data on establishments provided by the Institute for Employment Research. First, we present the two data sources and the identifiers that can be used to perform record linkage. Next, we describe our approaches to data preprocessing and matching algorithms. The final sections report on the linkage results and discuss linkage quality and overall matching success. Since our primary focus is on the specifics of the SOEP-LEE linkage project at hand, we do not provide detailed information on record linkage methods in general (see for example Herzog et al., 2007). A more detailed description of the record linkage procedures applied by the German Record Linkage Center can be found in Schild and Antoni (2014) or Gramlich (2014).

2 Data Sources

2.1 Survey data

The present record linkage project aims at matching administrative records to the establishments of the SOEP-LEE survey data (Liebig and Schupp, 2014). The purpose of the SOEP-LEE project was to create a linked-employer-employee (LEE) dataset which combines information on employees from the German Socio-economic panel study (SOEP) with data on their employers (Weinhardt et al., 2016). In order to achieve this, all SOEP respondents who were employed in 2011 were asked to provide the name and address of their employer at the moment of the SOEP interview. This was done retrospectively and has to be kept in mind, as this may have consequences for the quality of the address data. While this method to collect address information on establishments is error prone to some extent, it also has to be considered that these addresses were used to survey establishments. For the establishment survey, the address data was cleaned and validated before fieldwork. Only establishments which gave consent to the linkage are used for this linkage project. This means that an establishment was actually identified and interviewed under this address information.

Overall, 1708 establishments took part in the study. The consent rate for linking data was 34.4 percent. The address file of the SOEP-LEE survey thus contains 587 entries of establishments that took part in the survey and agreed to a linkage of their survey data with administrative records of the Institute for Employment Research. The following fields are included in the data set and could be used to link data records:

- Name of establishment – including legal form (where applicable)
- Postal code
- City name
- Street name and house number (combined field)

2.2 Administrative data

Administrative data on establishments originate from employment notifications collected by German social security agencies since 1975. An anonymous research data set on establishments, the Establishment History Panel (BHP), is offered by the Research Data Centre of the Federal Employment Agency at the Institute for Employment Research (see the BHP data report by Gruhl et al., 2012). Since the research data are anonymized and do not contain any name or address information, the linkage of records is conducted using information drawn from the establishment file of the Statistical Division of the Federal Employment Agency. The following identifiers were available:

- Name of establishment (partially including legal form)
- Postal code
- City name
- Street name and house number (combined field)

Since the reference period of the SOEP-LEE survey is 2011[1], the administrative address file was restricted to the amount of address notifications that were valid during the year of 2011. In other words, all addresses were selected that were valid at one point during that year, but not necessarily throughout the whole year. In some cases, new entries are created for existing establishments when small corrections are made. Most new entries reflect only some minor corrections of spellings or the name of the establishment. Sometimes though, establishments are given a new establishment ID in order to reflect a change in economic activity, ownership, or a relocation. All available entries in the 2011 file are selected and used for record linkage. A successful link to the administrative research data set however will only be achieved for those establishments that meet the sampling criteria of the Establishment History Panel (see section 5.2).

It is important to note that the administrative data are based on employment notifications of employees. The address is supposed to be that of the locality where the respective employee actually works, not the company headquarters or human resource department etc. The units covered in the administrative data are establishments. An establishment is defined as a commercial unit that is distinct regarding locality (within a municipality / community) and field of economic activity (according to the German classification of Economic Activities). In the survey data the definition of an establishment was implemented likewise.

3 Preprocessing procedures

The German Record Linkage Center uses a variety of preprocessing scripts in order to cleanse and standardize name and address data. These routines are described in more detail by Schild and Antoni (2014).

[1] The field period of the SOEP-LEE survey was from August 2012 until March 2013, but for several reasons the survey questioned the characteristics of the establishment in the year 2011.

The basic rules to modify string variables (such as establishment name, street name, and city) include:

- Replacement of German umlauts and other non-ASCII characters with ASCII equivalents
- All characters set to uppercase
- Removal of leading and trailing blanks
- Removal of punctuation characters

Apart from those basic modifications, several variable-specific routines developed at the Institute for Employment Research were applied to parse and standardize the fields used for record linkage (see Schäffler, 2014). As to the establishments' names, in both the survey and the administrative address data file there was a field containing establishment names which partially included legal forms (see section 2). The legal form was not always included, with the missing share (i.e., the number of observations for which no legal form could be extracted) being particularly high in the survey address file (78%). The legal form appears to have been given mainly in those cases where it was part of the popular firm name. The German Record Linkage Center's routines extract the legal form based on string pattern recognition procedures and create separate variables for name and legal form. After extracting the legal form, the name is standardized according to the steps listed above and other more variable-specific parsing rules.

Street names and house numbers were extracted from a combined address field using pattern recognition routines based on regular expressions. Street names were standardized and some common patterns were corrected, including resolving common abbrevations and replacing numbered street names with literal string components. The German word "STRASSE" (*street*) was consistently shortened to its usual abbreviation "STR". House numbers were processed to contain only numeric characters, all supplementary information is discarded.

The preprocessing of city names comprised the correction of common spelling mistakes and the solution of abbreviations. Common suffixes to city names were harmonized (such as in the case of *Frankfurt am Main* or *Berlin-Kreuzberg*).

Since we had access to the raw address material and the number of observations in the survey address data was rather limited, we had the possibility to perform manual data cleansing in individual cases where the first results of the linkage procedures suggested that linkage failed because of obvious errors in the data. These manual edits (of the survey address data) included:

- Common or obvious abbreviations of establishment names are resolved.
- Small modifications (regarding space and punctuation characters) are made to the establishment name in order to allow for automatic recognition and extraction of the legal form.

- In some cases, district names were given in the city field instead of the name of the superordinate municipality, along with the correct zip code. This seems to be an idiosyncracy of German addresses. However, since the zip code is known in these cases, the correct city name could be easily looked up.

4 Linkage strategy

4.1 Linkage methods

After applying data cleansing and parsing routines, we use the standardized names and addresses to identify matches between survey and administrative records.

The linkage of records is trivial in those cases where (preprocessed) identifiers perfectly correspond between the two data sources. In those cases where there is an exact accordance of all relevant identifiers (establishments names, legal forms, and addresses), we matched the records and labeled the match as *exact*. Note however that this step is performed on the standardized data, meaning there is a perfect accordance between the standardized data that underwent some basic cleaning and correcting steps (as described above) and therefore there were some steps that could at least in theory be prone to human error. Exact matches also include a few cases where one or two deviations on certain fields were tolerated: A non-conformity of the variable *legal form* was considered tolerable since that field is often missing. Also, we allowed for deviations regarding the house number **or** the zip code, because both are numeric fields and are prone to input errors or false remembering. A deviation on both the legal form **and one of** the fields *house number* or *zip code* was also permitted but occurred in just a few single cases.

Since perfect accordance of all relevant identifiers is rather the exception in real-world data, the error-proneness of address records renders the usage of inexact string comparison necessary. Especially when matching on string fields such as establishment name, city name, or street name, we need to allow for small deviations within each field because these variables do contain a lot of variation such as typos, misspellings, use of abbreviations, or different orderings of name components. We therefore perform distance-based record linkage to match observations that exceed a certain threshold on a quality variable composed of an index value of string similarity measures. In so doing, we employ commonly-used measures of string similarity such as N-grams and the Jaro-Winkler algorithm.

We use the Merge ToolBox (MTB) software (Schnell et al., 2004) to perform distance-based record linkage of the two address data files. The fields *street* and *city name* were compared with N-grams throughout. For establishment names, we used the Jaro-Winkler algorithm instead of N-grams in some matching runs because it assigns more weight to an agreement at the beginning of a string. This generated some additional matches in those cases where there is some suffix to the establishment name but a common beginning. The initial matching runs were based on all fields, but in order to allow for small deviations, some records were matched in omission of **one of** the fields *legal form*, *house number*, or *zip code*.

The following listing summarizes the linkage steps:

1. Exact Matching on *Name of establishment*, *Legal form*, *Street*, *House number*, *Zip code*, *City*
   (Partially disregarding *House number*, *Legal form*, *Zip code*)

2. Deterministic (distance-based) matching
   - *Name of establishment*: N-Gram match or Jaro Winkler match
   - *Street*, *City name*: N-Gram match
   - *House number*, *Zip code*: Exact matching

The distance-based matching runs generate quality indexes reflecting the cumulative string similarity of all fields used. Those links exceeding a specified threshold on the quality variable are considered matches. With a threshold of 0.95 per field for N-Gram similarity and 0.85 for Jaro-Winkler, the quality thresholds were set at rather high levels in order to limit the extent of false positives (i.e., non-matches spuriously classified as matches). This restriction seems to be appropriate given that there are only two fields available (name of establishment and legal form) to discriminate between units with the same or a similar address. Yet, after completing the matching steps described above, some manual review of the data was necessary to link records where the quality index did not exceed the classification threshold.

Generally, when linking data on firms, matching on address fields may result in a bias of the linked sample towards small enterprises that have one address only, thus systematically under-representing larger companies that have multiple plants with different addresses (or a larger building complex spanning over more than one street). Also, employers should report the actual workplace (i.e. plant) of their employees in social security notifications rather than the company headquarters or cost centers etc., but some employers deviate from that rule and therefore confound the assignment of workers to establishments. In the present linkage project, the surveyed establishments are the SOEP persons' workplaces and the survey data thus only apply to the establishment where the interview took place. We therefore decided to include addresses as matching variables. In those cases where employment notifications contain the address of a different workplace within the company, it is not possible to accurately link administrative information on the surveyed establishment. A manual review of the non-matched survey establishments suggests that there is no significant bias toward small businesses in the linked data set.

4.2 Multiple matches

The survey address file is unique with regard to establishments. In the administrative data on the other hand, multiple entries per establishment are possible. In some cases, establishments give the social security notifications for their employees with different establishment IDs reflecting a diversity of functional fields within the firm. As a result, there may be more than one valid match per survey establishment. From the address material, there is no telling what the correct match is, or which is the closest correspondence to the unit that

was questioned in the survey. Another explanation for the presence of multiple matches is, as described in section 2.2, that new entries are created when making small modifications regarding name or address of the establishment. For those reasons, all plausible matches are kept at that stage. The invalid ones however are filtered out when merging the BHP data.

4.3 Blocking

To limit the number of comparisons necessary and thereby speed up computation time, all matching procedures using string similarity were performed with a blocking on the first 3 digits of the zip code. Blocking means to use a relatively coarse variable to form subsets of the data and to compare only those observations that share this common field. The first 3 digits of the zip code are a fairly convenient blocking variable because the field usually has a good quality (with errors mainly affecting the two last digits of the zip code which vary within cities), is seldom missing, and it forms relatively clear-cut geographical areas. In our case, comparisons were made in a total of 318 groups formed by the first three digits of the zip code.

5 Linkage results

5.1 Matches between survey and administrative address files

Of the 587 establishments taking part in the SOEP-LEE survey and consenting to data linkage, there was an exact match on all relevant fields for 61 establishments (see table 1). Another 35 establishments could be linked based on an exact agreement on **all but one** of the fields legal form, zip code, and house number. To be more precise, this number includes a few single cases with two deviations, where both the legal form and either the house number or the zip code did not coincide.

The large majority of matches were accomplished in the distance-based matching runs. This results from the simple fact that an exact conformity of firm names is rarely given in two address data files with very different data collection processes. This issue will be further discussed in section 6. Additional matches could be achieved by leaving out **one of** the error-prone address fields (*legal form*, *house number*, *zip code*).

Since we used rather restrictive thresholds in order to limit the number of incorrectly linked records, some establishments could not be linked by automatic matching routines. Here, the variations between the two data sources were either too big or the name had too little discriminatory power to reach the classification threshold. In these cases, a manual review of all possible links was performed in order to identify valid matches. This manual review added matches for another 77 establishments.

Note that table 1 does not include multiple matches in the administrative data. The numbers reflect the best matches per surveyed establishment only.

Table 1: Matching results

Matching algorithm	Survey establishments matched
	N
	% of total survey file
Exact matches	61
	10.39
Exact matches (disregarding 1-2 fields)	35
	5.96
Distance-based matches	196
	33.39
Distance-based matches (disregarding 1 field)	116
	19.76
Manual matches	77
	13.12
Total	**485**
	82.62
Total survey file	587
	100.00
Unmatched	102
	17.38

Note: In case of multiple matches per survey establishment, the numbers given in the table reflect the best match only.

5.2 Link to Establishment History Panel

The final record linkage product is an anonymized data set containing numeric and non-systematic ID variables as well as a measure of the respective matching quality. It does not entail any sensitive information on names or addresses. The linkage data set can be used to link the survey data with a sample of the administrative Establishment History Panel (BHP).

The Federal Employment Agency address data file covers any firm reporting employees. The sample of the Establishment history panel however consists of establishments with at least one employee liable to social security or in marginal employment at the cutoff date of June 30th of each year. Due to this restriction, a certain number of small establishments that do not meet this criterion (e.g., small firms with only short-term employees) show up in the address data file but are not part of the Establishment History Panel. The following table gives an overview of the actual number of observations that could be linked to the Establishment History Panel at the cutoff date of June 30th 2011. The number given in parentheses includes multiple matches per surveyed establishment.

For each observation in the linkage data set, the *grade* variable reflects the quality of the match. Where there are multiple administrative data matches per survey establishment, users can sort observations in descending order of matching quality by using the following Stata command:

```
bysort pnrfest (grade):  gen n = _n
sort pnrfest n
```

Table 2: Number of matched establishments in 2011

Matching algorithm	Establishments in linked dataset	
	N	(N mult.)
Exact matches	59	(60)
Exact matches (discarding 1-2 fields)	34	(38)
Distance-based matches	181	(218)
Distance-based matches (discarding 1 field)	115	(197)
Manual matches	59	(59)
Total	**448**	(572)

6 Discussion

The present report provides an overview of the record linkage of the SOEP-LEE survey data and administrative data on establishments of the Institute for Employment Research. It includes an overview of the two data sources, the preprocessing methods and matching algorithms used to link the records. The matching was conducted in order to provide a linked data set combining administrative records and survey data on establishments. With a matching rate of 82.6 percent, the record linkage proved successful.

Record linkage of data on establishments is, at least in our experience, usually more demanding than the linkage of data on persons. The latter usually includes a slightly larger number of identifying characteristics (first and last name, in some cases birth name or further name components, and birth date). What is more, names of establishments are more likely to vary between data sources: Name components could be mixed up and there might be variations to a firm name or some supplementary information (e.g., department, subtitle) given in the one data source but missing in the other. Also, the use of abbreviations is a lot more common than in data on persons, since interviewees are more prone to use short versions of establishment names in surveys than they are to reporting nicknames instead of their full civil name. Another issue is the variation of establishment names over time (e.g., take-overs, change of legal form).

Another issue is the ambiguousness of establishment units. Where there are several establishments, organizationally linked, at the same location, it is not always clear which observation is the true match by just looking at the address data. In the present case of the linkage of the SOEP-LEE survey data with administrative establishment data of the Institute for Employment Research, we provide all possible links that have passed the above-mentioned quality thresholds. Therefore, we leave room for researchers using the linked data set to use the survey data to single out the corresponding units where multiple options were given.

Especially when linking establishments, automatic matching routines yield limited results. In the present linkage project, the matching rate could be substantially increased by means of manual or script-based preprocessing of both data sources on the one hand and by manual record linkage on the other. In general, the full potential of automatic record linkage could be exploited by collecting as much information as possible. Ideally, full establishment names should be recorded during data collection, but variations of the establishment name,

i.e. common abbreviations or alternations, should be enquired as well. Furthermore, an establishment's legal form is of high value when identifying establishments that belong to a larger network of firms and should thus be asked for in the survey by all means.

References

Gramlich, T. (2014). *'STROKES' – Record Linkage der Schlaganfälle in Hessen 2007-2010*. German RLC Working Paper No. wp-grlc-2014-03.

Gruhl, A., A. Schmucker, and S. Seth (2012). *The Establishment History Panel 1975-2010. Handbook version 2.2.1*. FDZ-Datenreport 04/2012 (en).

Herzog, T. N., F. J. Scheuren, and W. E. Winkler (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Liebig, S. and J. Schupp (2014). *SOEP-LEE Betriebsbefragung - Die Betriebsbefragung des Sozio-oekonomischen Panels. Datenzugang über DSZ-BO. DOI:10.7478/s0549.1.v1*.

Schäffler, J. (2014). *ReLOC linkage: a new method for linking firm-level data with the establishment-level data of the IAB*. FDZ-Methodenreport 05/2014 (en).

Schild, C. J. and M. Antoni (2014). *Linking Survey Data with Administrative Social Security Data - the Project "Interactions Between Capabilities in Work and Private Life"*. German RLC Working Paper No. wp-grlc-2014-02.

Schnell, R., T. Bachteler, and S. Bender (2004). "A Toolbox for Record Linkage". In: *Austrian Journal of Statistics* 33.1-2, pp. 125–133.

Weinhardt, M., A. Meyermann, S. Liebig, and J. Schupp (2016). *The Linked Employer–Employee Study of the Socio-Economic Panel (SOEP-LEE): Project Report*. SOEPpapers No. 829.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 100
D-90478 Nuremberg

Editors

Rainer Schnell, Manfred Antoni

Template layout

Christine Weidmann

Download

www.record-linkage.de

The German Record Linkage Center was funded by the German Research Foundation (DFG).